Cluster Analysis of Stocks
Project Proposal Paper

Jordan Zane Caleb Clough Matt Dyer Evan Linden

PROBLEM STATEMENT

This project will aim to mine extensive longitudinal American stock/etf market data to identify interesting trends.

In particular, our group aims to cluster individual stocks into groups to see what individual stocks tend to "move" together. Ideally this analysis will uncover greater nuance and cross relationships outside of typical stock groupings (technology, manufacturing, finance, etc.) and allow for novel (or atleast modified) classification grouping schemes.

Potential applications of the results of this project could be to more accurately predict stock movements, or atleast correlations between certains stocks/etf that are found to be in similar groups. Not only could this data be useful for identifying undervalued assets, but it could also be useful to hedge risk in a portfolio by ensuring multiple cluster groups are represented in a portfolio-- ensuring groups tend to move independently, while having multiple groups present in a portfolio can ensure the overall portfolio balance does not swing wildly.

DATA SETS

Our main dataset is a .csv containing 5 years of AMEX, NYSE, and NASDAQ end of day data. The source of this data is Kaggle and the URL is as follows:
found at:
https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

https://www.kaggle.com/qks1lver/amex-nyse-nasdaq-stock-histories?select=fh_5yrs.csv

The overall size of the dataset is 712 MB, with a total of 6852039 data points. The data is structured such that each stock ticker represented has (up to) 5 years of daily data.

Each data point has a total of 8 attributes. They are as follows, with information about the type of variable each is:

- Date - Ordinal
- Volume - Numeric (Ratio)
- Open - Numeric (Ratio)
- High - Numeric (Ratio)
- Low - Numeric (Ratio)
- Close - Numeric (Ratio)
- Adj Close - Numeric (Ratio)
- Symbol - Categorical

The data appears to be very clean, but likely some minor cleaning will need to be done, as described in the aforementioned sections.

Additionally, we have a secondary data set to supplement the main dataset in the event that the primary dataset is not sufficient to meet our project objectives. This secondary dataset can be
This dataset is structured a bit differently than the primary dataset. Each stock and etf has a separate txt file with longitudinal data. Each txt file is structured with the data being comma
delimited, and thus can be treated as a csv with minimal modification. In total there are

8539 files totaling 771 MB. Each file contains a total of 7 attributes. They are as follows, with information about the type of variable each is:

- Date - Ordinal
- Open - Numeric (Ratio)
- High - Numeric (Ratio)
- Low - Numeric (Ratio)
- Close - Numeric (Ratio)
- Volume - Numeric (Ratio)
- Open Int - Numeric (Ratio)

This data also appears to be very clean, but may require substantial post-processing to integrate with the primary dataset in the event that it is needed

LITERATURE SURVEY

Studies come from Kaggle:

- Which stocks should I invest in? - included daily percentage change in price and volume, Correlation of percent change, Predicting expected gain and loss for each ticker, and growth trends of tickers
- Stock diversity analysis - analysing clusters to identify stocks with opposite trends
- Linear regression of stock histories - using a regression model to predict stock prices
- Wsb stock analysis - using popular tickers mentioned on reddit.com to analyze growth

PROPOSED WORK

1 Data Preprocessing

Once we have collected our complete data set from Kaggle, we will need to clean the data before we do any analysis with it. The data cleaning process will begin by dropping any NULL values from the dataset. Then the next step is to replace any string values such as '$' with a ' ', so we do not have any strings in our data set. Considering we are going to be comparing the stocks based on their prices over time, we will also need to convert the dates using the Pandas function "to_datetime", so that Python recognizes them as Dates, not strings. The last step of data cleansing is to locate and remove any outliers from the data set. Once these steps have been completed, the dataset will have been cleansed of any imperfections and is now primed for any data-mining we need to pursue.

For data integration, we will be using potentially multiple stock datasets containing similar attributes. To avoid redundancy in the data, the standard stock symbols will be used for entity identification. The dates will be converted to datetime using the Pandas library in Python to ensure that all dates are in a standard tractable format. This will ensure that no duplicate dates will be used for the same stock.

2 Data Analysis

Our data analysis will rely on hierarchical clustering which requires high space and time complexity. Therefore, we will need to use considerable data reduction to ensure efficient and speedy mining techniques. Since we are primarily interested in the daily stock movement, we can reduce the dimensionality of the data by only using the open and close attributes to track daily stock movement. This will also allow us to avoid possible strong correlation in attributes such as close and adjusted close. In case the data needs to be further reduced for effective hierarchical clustering, we can turn to numerosity reduction

techniques like sampling. For our time series analysis, we will reduce each cluster to an ARIMA regression model to forecast future stock movement.

3 Literature Survey Comparison

First off there is only one piece of work that is related to analysis using clusters, the rest are analysing the data set using methods such as correlation of percentage change and linear regression. The one piece of prior work, “Stock Diversity Analysis”, analyzes with percentage change of day-to-day price changes which will be similar to ours, but what we will be doing differently is using hierarchical clustering. Using the ARIMA regression model we will hopefully be able to get more accurate and better predictions.

DATA ANALYSIS AND TOOLS

First we will measure daily stock movement by subtracting the open price from the closing price of each stock for every day. Next, we will randomly select 80% of the days to be our training set for the model with the remaining 20% serving as our test set. We will then use agglomerative hierarchical clustering to determine similar groups of stocks according to their daily price movements.

This will be performed using the AgglomerativeClustering library in Python. Euclidean distance will be used as the similarity measure and mean distance will be the metric to determine the distance between clusters. These metrics are purely arbitrary and we may experiment with different distance metrics to assess the differences in clustering. Using a dendogram to represent the process, we will select a termination condition that agrees with our domain knowledge. For example, if one cluster contains mainly tech stocks and energy sector stocks, we would continue the partitions until these groups of stocks are largely segregated. This will be a difficult judgment and require considerable domain research to determine when to stop clustering. After all, our goal is to find interesting and less obvious groups. Therefore, we want to avoid overfitting the data.

Finally, we will predict future movement of the means of these clusters using time series analysis in R. In particular, we will use the Hyndman-Khandakar algorithm. Using plots of the ACF of the residuals in conjunction with portmanteau residual tests, we will adjust the model accordingly until the residuals appear as white noise. Then, we will use the resulting models for each cluster to forecast future stock movement.

EVALUATION METHODS AND TOOLS

For our purposes, we are assuming the ground truth is essentially unavailable because we are looking for clusters of seemingly disconnected stocks. Therefore, we will rely on intrinsic methods to evaluate how well the clusters are separated and how compact the clusters are.

For a general guideline, we will use the elbow method to determine if the optimal number of clusters were used. However, we will still ultimately rely on our own judgment and domain expertise to decide on the final number of clusters. We will ensure that our

judgment generally agrees with the elbow method.

Our primary evaluation method will be through cross-validation. We will test our models that use a different number of clusters, different distance measures, etc. and evaluate the sum of the squared differences in between the points in a cluster and their centroid as well as the squared differences with the other centroids. A smaller intradistance and a larger interdistance signifies a better fitting model.

MILESTONES

1. Locate Data Set(s) - Completed
2. Data Preprocessing - 7/12-7/18
3. Data Analysis - 7/19-7/25
    a. Daily Stock Movement
    b. Select Training and Test Sets
    c. Clustering code
    d. Analyzing data
4. Making a conclusion from our analysis - 7/27-8/1